The CDF Run 2 Computer Farms

Stephen Wolbers

Fermilab

For the CDF Farms Group and the Fermilab Computing Division Farms Group

September 4, 2001

Outline

- **Introduction to Run 2 Data Rates/Processing Needs**
- **Architecture of the CDF Run 2 Farms**
- **Experience with the farms in Run2**
- **Future**
  - Run 2a
  - Run 2b

Introduction

- CDF Run 2 Data Rates are substantially larger than Run 1 (factor 20 higher).
  - 20 Mbyte/sec to tape peak
  - Approximate 250 Tbyte/year to tape
- This data must be processed as quickly as it is collected (with a short time delay for preparing calibrations).
- The output data has to be organized into well-defined physics data sets.
- In addition, reprocessing and simulation are also required.

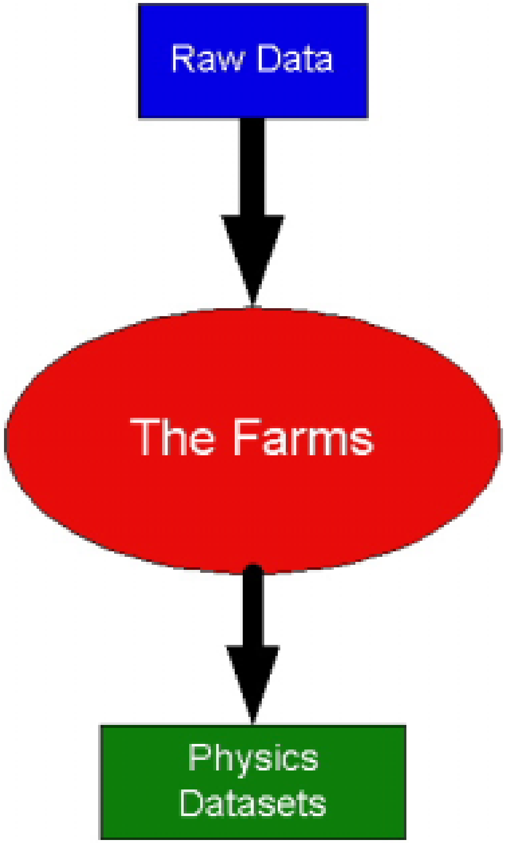
CDF Run 2a Farm Computing

- **Goal: CPU for event reconstruction of about 5 sec/event on a PIII/500 MHz PC (Each event is 250 Kbyte).**
- **Assuming 20 Mbyte/sec peak (approx. 75 Hz)**
  - Requires 375 PIII/500 processors to keep up
  - Faster machines -> Fewer processors required
    - So 180 PIII/500 duals will suffice.
    - Or 90 PIII/1 GHz duals.
- **Requirement is reduced by accelerator/detector efficiency and increased by farms inefficiency.**

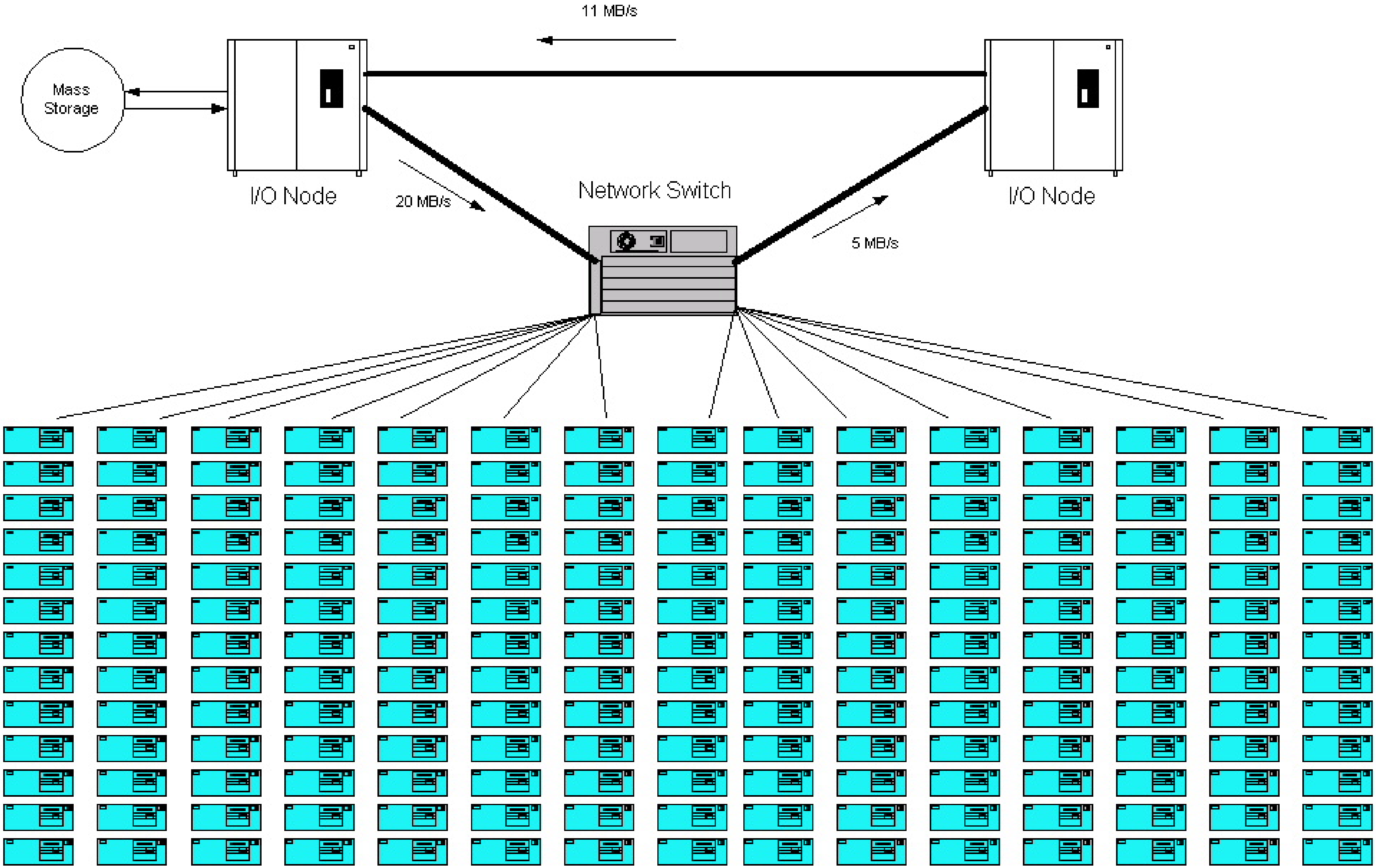
CDF Offline Production Farms for event reconstruction

- The CDF farms must have sufficient capacity for Run 2 Raw Data Reconstruction.
- The farms also must provide capacity for any reprocessing needs.
- Farms must be easy to configure and run.
- The bookkeeping must be clear and easy to use
- Error handling must be excellent.

Simple Model



Run I I CDF PC Farm



Design/Model

- **Hardware**
  - Choose the most cost-effective CPU's for the compute-intensive computing.
  - This is currently the dual-Pentium architecture
  - Network is fast and gigabit ethernet, with all machines being connected to a single or at most two large switches.
  - A large I/O system to handle the buffering of data to/from mass storage and to provide a place to split the data into physics datasets.

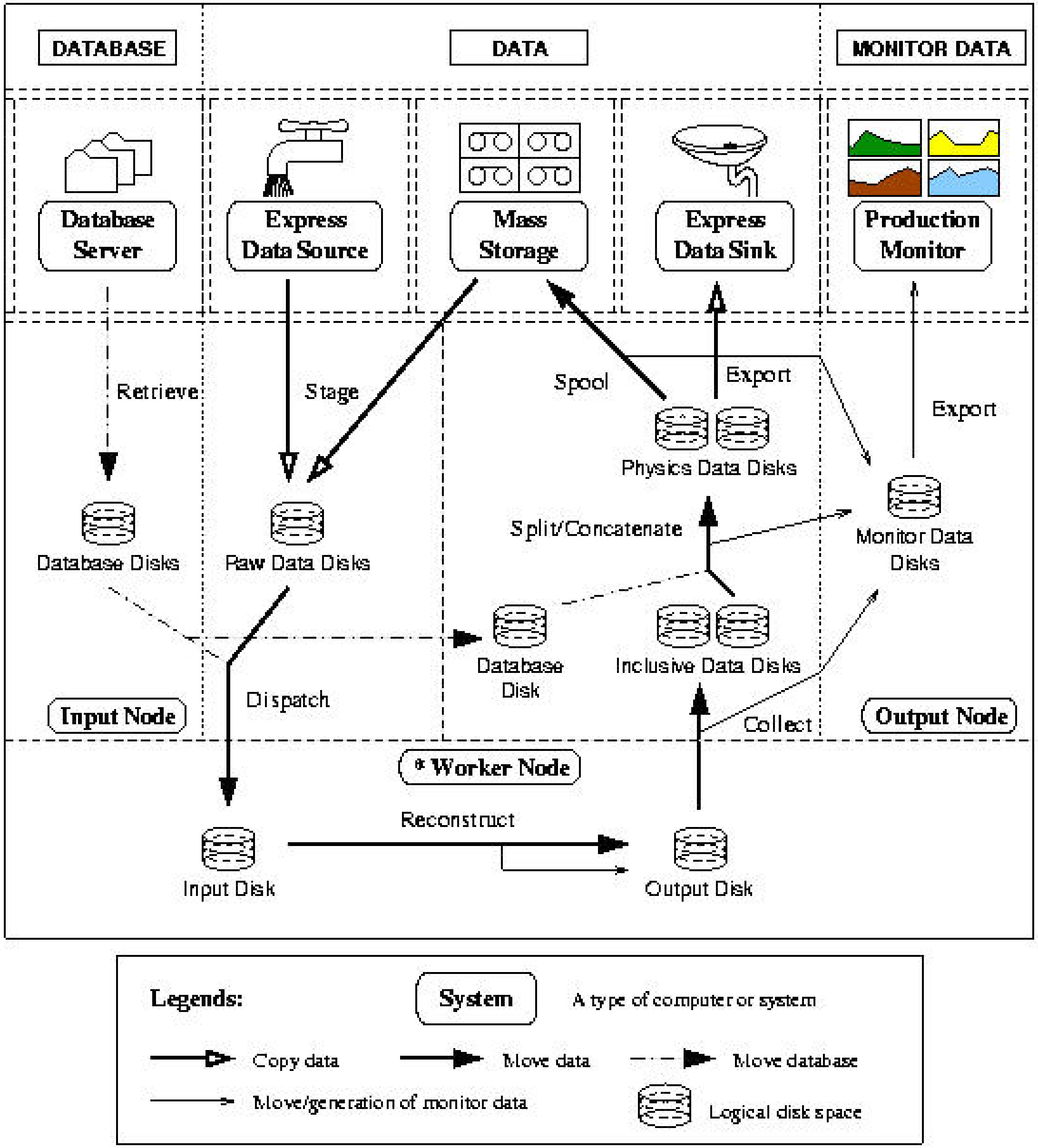
EXIT

1/24/2001

Software Model

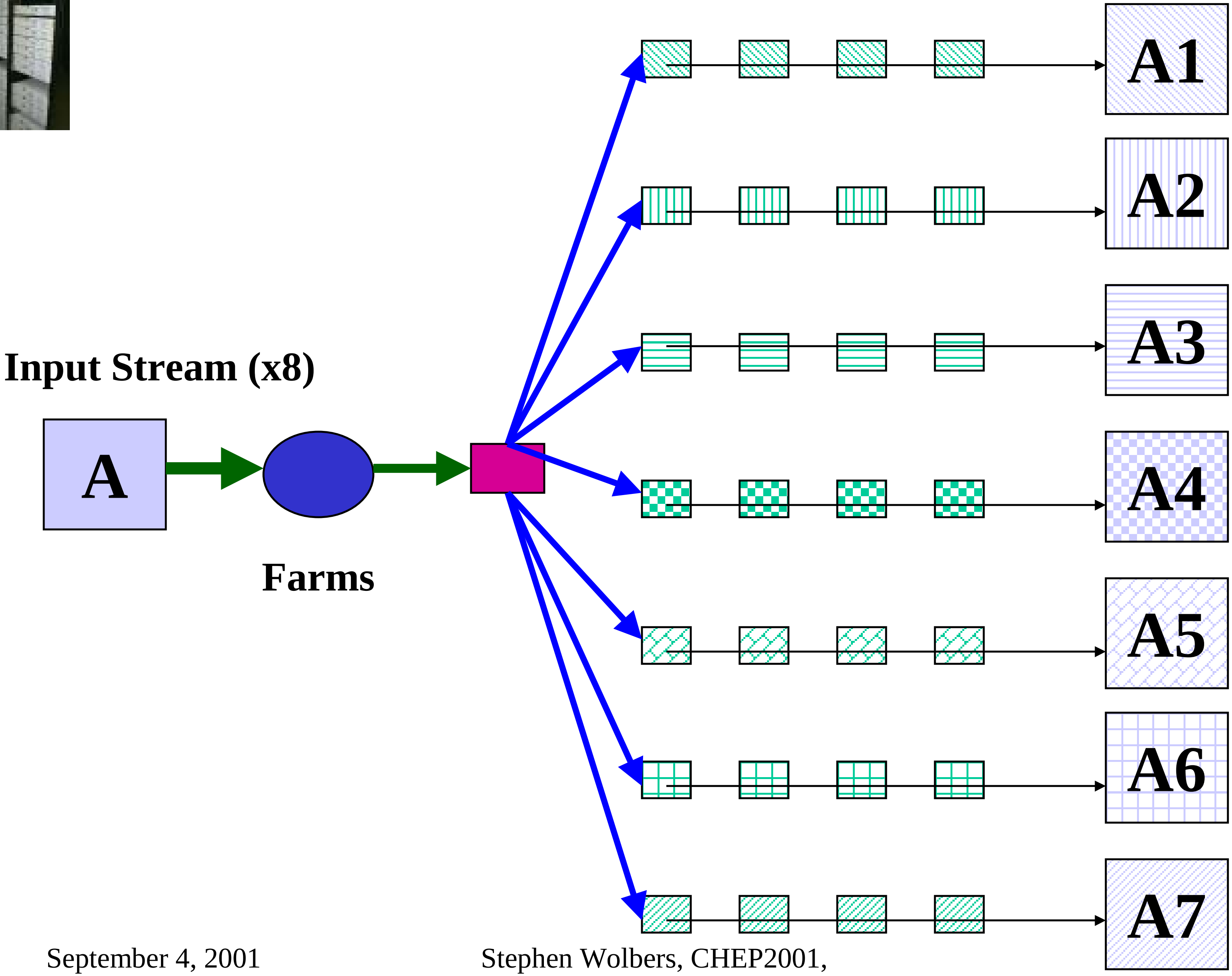
- **Software consists of independent modules**
  - Well defined interfaces
  - Common bookkeeping
  - Standardized error handling
- **Choices**
  - Python
  - MySQL database (internal database)
  - FBSNG (Farms Batch System)
  - FIPC (Farms Interprocessor Communication)
  - CDF Data Handling Software

Conceptual Model of Run 2 Production System



Physics Analysis Requirements and Impact

- **Raw Data Files come in ~8 flavors, or streams**
  - 1 Gbyte input files
- **Reconstruction produces inclusive summary files**
  - 250 Mbyte output files
- **Output Files must be split into ~8 physics datasets per input stream**
  - Target 1 Gbyte files
  - About 20% overlap
- **Leads to a complicated splitting/concatenation problem, as input and output streams range from tiny (<few percent) to quite large (10's of percent)**



Input Stream (x8)

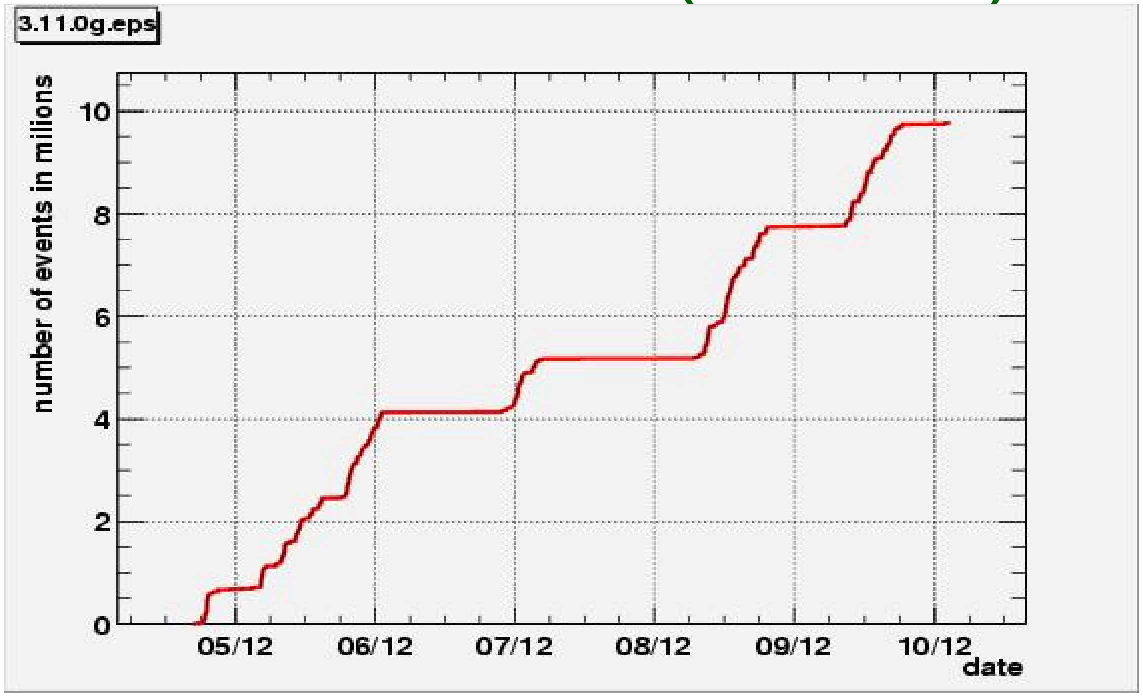
Status of CDF Farms Hardware

- **154 PC's are in place.**
  - 50 PIII/500 duals
  - 40 PIII/800 duals
  - 64 PIII/1 GHz duals
- **I/O nodes are ready (more disk is being added for output buffering).**
- **The CDF Data Handling System has sufficient capacity to handle the I/O to/from tape.**

Experiences so far in Run 2

Early Processing Experience

- **Commissioning Run (October, 2000)**
  - Ran 4 weeks after data was collected.
  - 9.8 Million Events, 730 GB input, 1080 GB output.
  - CPU/event = 1.2 seconds (PIII/500)

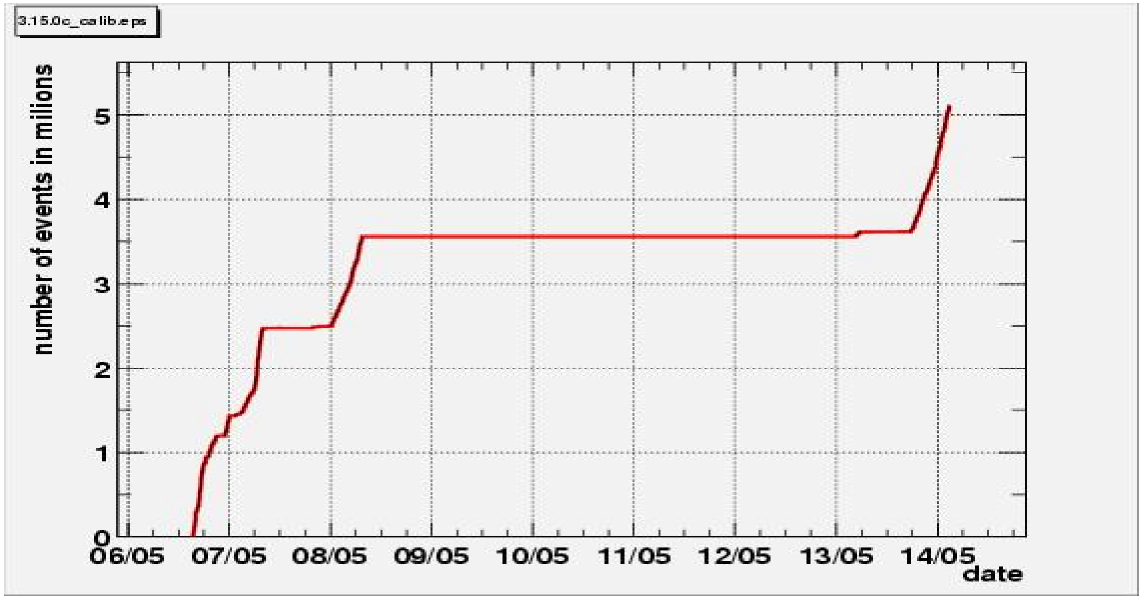


Lessons from Commissioning Run

- **Data size was not an issue. Farms could easily keep up.**
- **I/O was problematic. It was easy to flood the system, fill disk buffers, etc.**
- **Reconstruction code was an issue. Modifications were common, leading to occasional delays.**

Early Processing Experience

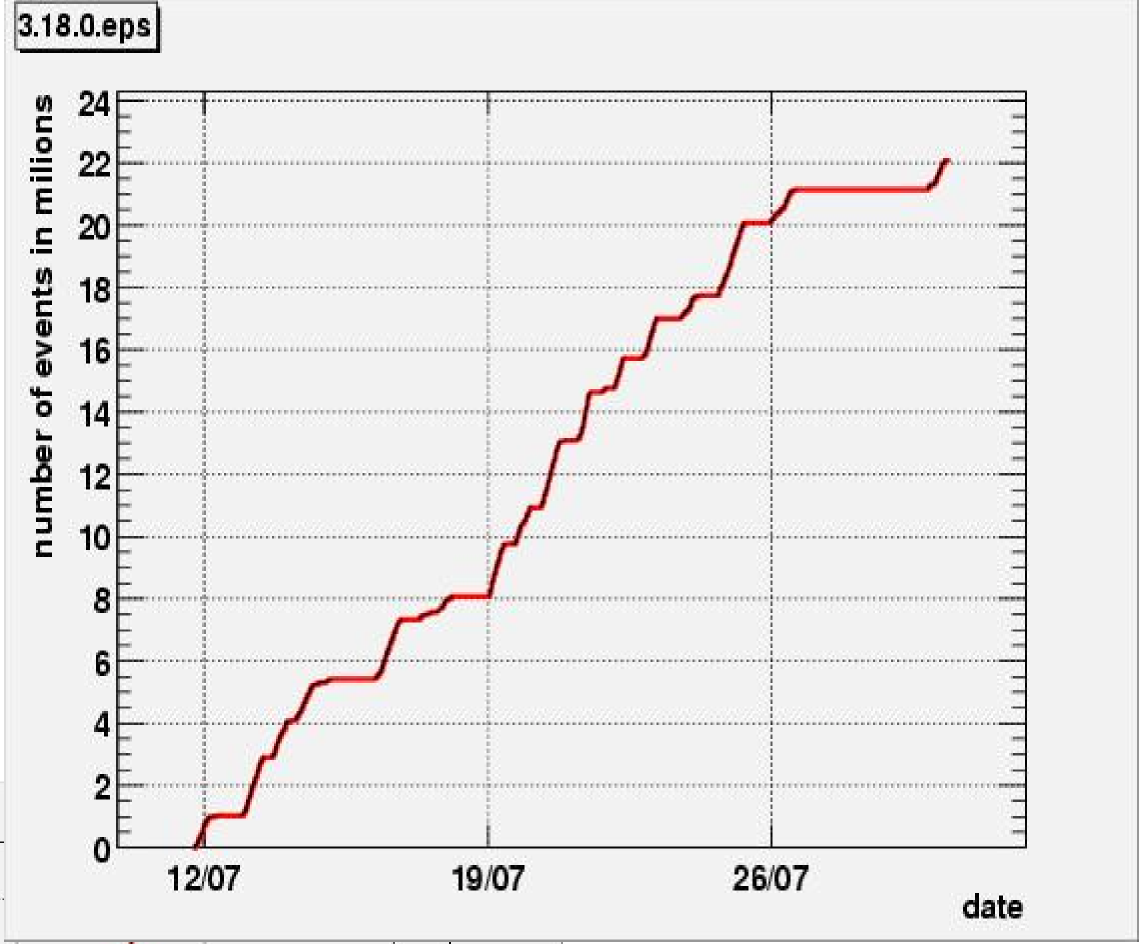
- **April 2001 Data**
  - First 36x36 bunch collisions.
  - Ran about 1 week after data was taken.
  - 5.1 Million Events, 1.2 TB input, 1.6 TB output
  - CPU/event = 1.0 seconds (PIII/500)

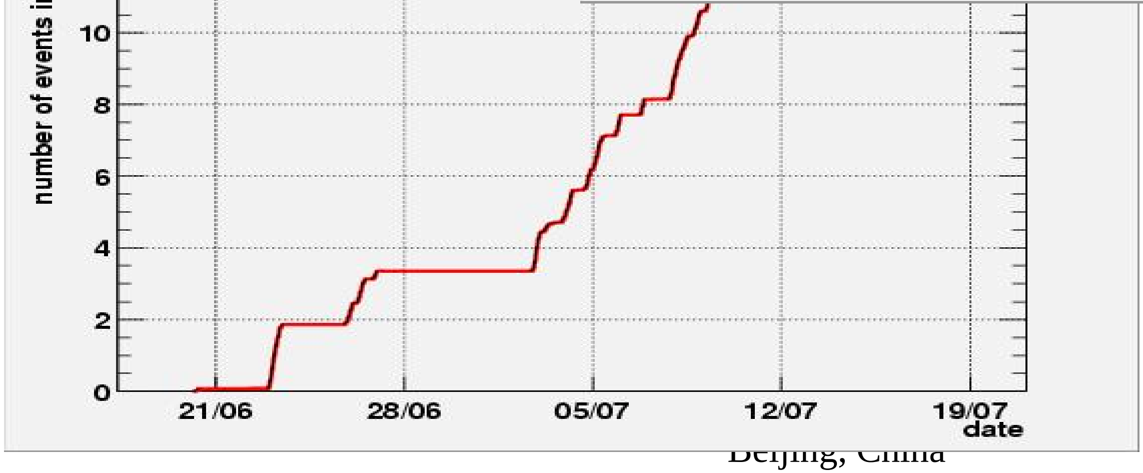


June-July 2001 Data

- **First substantial data taken in Run 2**
  - Approximately 34 million events (“good runs”).
    - Approximately 8.5 TB of data.
    - Approximately 10 TB of output data.
    - ~3 seconds/event on PIII/500
  - I/O system was still not fully operational at this time, and this led to a backlog of data.
  - Long accelerator downtime (unplanned) allowed the farms to catch up with the backlog of data.
  - Code modifications (mainly due to detector changes) were common.
  - Procedures for providing proper calibrations were tuned up during this time.
  - Full splitting into many output datasets was implemented.

June/July Processing





August-September Data

- **More data is being taken in August-September, 2001.**
- **The CDF detector is still changing, making calibrations and code changes more important.**
- **A “super-expressline” was invented to get data to physicists as quickly as possible.**
  - One stream (A) is processed as soon as the files are available.
  - These events are reprocessed later with final calibrations and code.

Run 2a Prospects

- **Run 2a will resume in November after the October shutdown.**
- **Luminosity is expected to increase.**
- **Data Handling system will be completed.**
  - This in turn will allow the farms to run at full rate.
- **The CDF Farms will be able to keep up with the data.**

Run 2b at Fermilab

- **Run 2b will start in 2004 and will increase the integrated luminosity to CDF and D0 by a factor of approximately 8 (or more if possible).**
- **It is likely that the computing required will increase by a similar factor, in order to pursue the physics topics of interest:**
  - B physics
  - Electroweak
  - Top
  - Higgs
  - Supersymmetry
  - QCD
  - Etc.

Run 2b Computing

- **Preliminary estimates for Run 2b computing:**
  - 8x CPU, disk, tape storage.
  - Expected cost is same as Run 2a because of increased price/performance of CPU, disk, tape.
  - Plans for R&D testing, upgrades/acquisitions will start next year.
- **Data-taking rate:**
  - Potentially 80 MB/s or more.
  - About 1 Petabyte/year to storage.

Summary

- CDF Production Farms are commissioned, tested and have processed tens of millions of events.
- Run2a will be a major task for the farms.
- Run2b is potentially substantially larger than Run2a, and some changes to the farms will likely be needed to address this.